Diff4TST: Masked Diffusion Language Model for Text Style Transfer

Xinchen Ma[1], Gaole He[2], Yunshi Lan[1]*, Weining Qian[1]
[1]Department of Data Science and Engineering, East China Normal University
[2]School of Computing, National University of Singapore
{xinchen.ma}@stu.ecnu.edu.cn, {hegaole}@nus.edu.sg, {yslan}@dase.ecnu.edu.cn

Abstract

Despite recent progress in LLMs for text style transfer, most existing methods rely on costly task-specific training and offer limited control over separating stylistic modification from content preservation. We propose Diff4TST, a diffusion-based language model that formulates text style transfer as an explicit copy-and-edit process. Built upon masked diffusion language models, Diff4TST introduces a style-aware noise schedule that selectively perturbs stylistic tokens while preserving content-bearing tokens during supervised fine-tuning. At inference time, we further introduce a generate-then-refine strategy that iteratively improves style compliance via gradient-based token remasking, without reinforcement learning or external reward models. Extensive experiments on both fine-grained and polarity-based benchmarks show that Diff4TST achieves substantially improved style accuracy and controllability while maintaining strong content preservation and fluency. These results suggest diffusion-based language models as a principled and effective alternative to autoregressive pipelines for text style transfer.

1 Introduction

Text Style Transfer (TST) aims to rewrite a given text to match a specified target style while preserving its original semantic content. (North et al., 2023; Zhang et al., 2015; Briakou et al., 2021). Most existing approaches adopt autoregressive Large Language Models (LLMs) to solve TST, treating style transfer as a left-to-right generation process. While effective, such methods typically rely on supervised fine-tuning or reinforcement learning to enforce stylistic constraints (Gong et al., 2019; Jin et al., 2022), thereby incurring substantial computational costs. Moreover, these pipelines often require carefully designed reward functions and multiple training stages, making them complex, unstable, and difficult to scale or deploy.

While autoregressive LLMs excel at open-ended text generation, they are not always well suited for text style transfer. In many style transfer scenarios, preserving the original semantics requires copying a large portion of the input text, while only a small subset of style-related tokens needs to be modified (Li et al., 2018a). Treating the task as full left-to-right generation forces models to repeatedly regenerate factual content (Guu et al., 2018), which can introduce unnecessary modeling complexity and dilute the learning signal for stylistic transformation, thereby reducing training efficiency. Inspired by recent progress in discrete diffusion language models (Nie et al.), we reformulate text style transfer as an explicit copy-and-edit process. Unlike autoregressive decoding, diffusion-based models generate tokens in parallel and support iterative refinement, naturally aligning with the localized and incremental nature of stylistic edits (Zhang et al., 2025). By enabling gradual, token-level modifications while preserving most of the original content, diffusion models provide a principled and efficient foundation for controllable text style transfer.

Despite this promise, applying discrete diffusion language models to controllable text style transfer remains underexplored. Existing diffusion-based approaches typically employ uniform noise schedules that mask tokens indiscriminately, encouraging global rewriting rather than targeted edits (Schiff et al., 2025). As a result, stylistically salient tokens are not explicitly prioritized during training. Moreover, the standard reverse diffusion process alone does not guarantee that generated outputs strictly satisfy desired style constraints.

To address these challenges, we propose **Diff4TST**, a masked diffusion language model tailored for arbitrary text style transfer. Built upon LLaDA (Nie et al.), Diff4TST introduces two key

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6736–6748

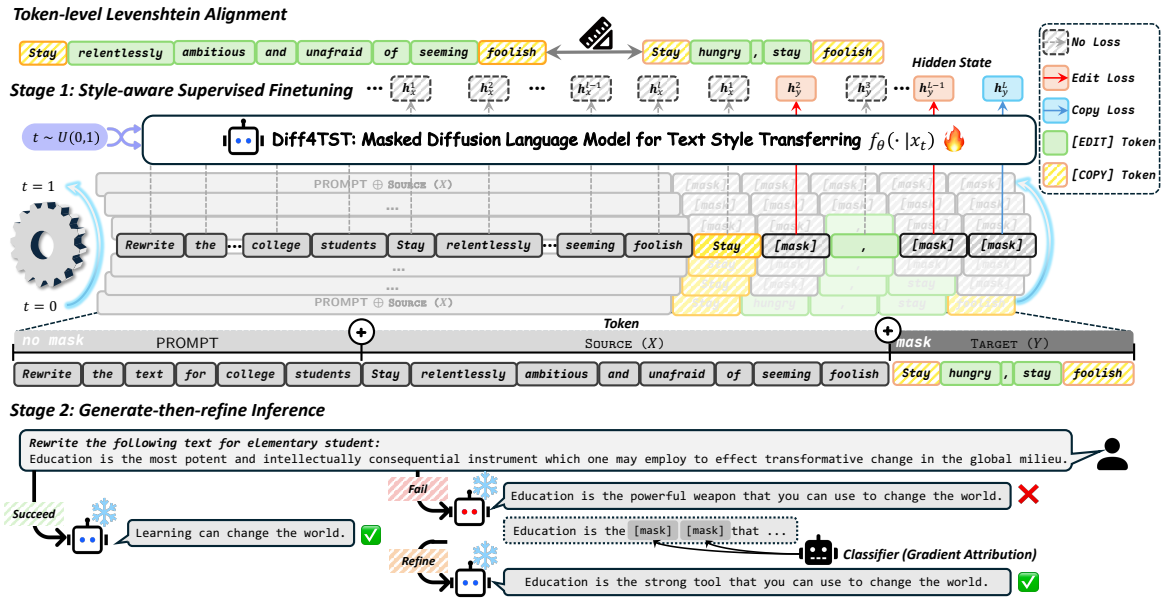

Figure 1: Overview of Diff4TST framework.

components. First, we design a *style-aware noise schedule* to involve in the supervised fine-tuning stage that explicitly distinguishes stylistic edits from content-preserving tokens, encouraging minimal and targeted rewriting. Second, we propose a *generate-then-refine* inference framework, which iteratively identifies and regenerates style-critical tokens using gradient-based attribution. This framework enforces style constraints without reinforcement learning or handcrafted reward models, resulting in a simpler, more stable, and easily deployable system. Extensive experiments demonstrate that Diff4TST generalizes well and achieves strong performance across diverse style transfer tasks, underscoring the suitability of diffusion-based language models for controllable text style transfer that requires precise and minimal edits. The contributions of this work can be summarized as:

- We propose Diff4TST, an innovative masked diffusion language model for text style transfer. This model involves a style-aware noise schedule that prioritizes stylistic pivots over uniform token corruption and a generate-then-refine inference framework that facilitates controllable rewriting.

- We conduct extensive experiments on four TST benchmarks, including polarity style and fine-grained style transfer. The new diffusion-based paradigm surpasses advanced baselines and achieves the best performance in the perspectives of accuracy, preservation, and fluency in most cases.

2 Related Work

Diffusion Models for Text. Diffusion models have shown strong performance in vision, especially for image synthesis and editing via iterative denoising (Ho et al., 2020; Rombach et al., 2022; Austin et al., 2021), motivating their extension to language generation. Early diffusion-based text models operated in continuous latent spaces, injecting Gaussian noise into token embeddings or latent representations (Hoogeboom et al., 2021; Li et al., 2022; Lin et al., 2023; Gong et al., 2023), but often relied on continuous relaxations or auxiliary autoencoders, limiting precise token-level control.

More recent work has explored discrete diffusion language models that directly operate on token sequences (Austin et al., 2021; Zhou et al., 2024). Mask-based discrete diffusion formulates generation as iterative denoising over masked tokens, enabling parallel prediction and bidirectional context modeling (Ye et al., 2025; Gong et al., 2025). This paradigm naturally supports localized editing and iterative refinement, making it a promising foundation for controllable text rewriting tasks.

Text Style Transfer. TST has been studied across diverse applications, including education, legal writing, and social media moderation (Tan et al., 2024a; Li et al., 2021). Existing approaches mainly fall into two categories: attribute substitution and

sentence generation.

Attribute substitution methods identify style-bearing tokens and replace them using lexicons, classifiers, or mapping functions (Li et al., 2018b; Sudhakar et al., 2019; Lee, 2020; Fu et al., 2018; Luo et al., 2019). While effective for coarse polarity transfer, they often suffer from semantic inconsistency and limited flexibility for fine-grained style control.

Sentence generation methods directly produce text in the target style using neural language models. Non-disentangled approaches, such as prompt-based rewriting and instruction tuning (Reif et al., 2022; Suzgun et al., 2022; Mukherjee and Dušek, 2023; Zong et al., 2024), are flexible but lack explicit mechanisms for regulating stylistic intensity. Reinforcement learning extensions (Deng et al., 2022; Liu et al., 2024) mainly optimize global style alignment. In contrast, disentangled methods (Han et al., 2024) introduce auxiliary objectives to separate style and content, enabling finer control at the cost of increased annotation and modeling complexity.

3 Preliminaries

3.1 Text Style Transfer

We assume a predefined style space $\mathcal{S}$, which represents a set of categorical or discrete stylistic attributes. Formally, given an input text X associated with a source style s_x, the goal of TST is to generate a target text Y that conforms to a specified style s_y, where $s_x, s_y \in \mathcal{S}$ and $s_x \neq s_y$. During training, paired examples $(X_i, Y_i)_{i=1}^{N}$ annotated with their corresponding styles are provided. At inference time, a TST system is expected to generalize to unseen inputs and transfer them to arbitrary target styles in $\mathcal{S}$, while maintaining the underlying meaning of the original text.

3.2 Masked Diffusion Language Models

Masked Diffusion Language Models (MDLMs) define a model distribution $f_\theta(x_0)$ through a forward process and a reverse process. x_0 is a sequence of tokens at $t = 0$. The forward process gradually masks tokens independently in x_0 until the sequence is fully masked at $t = 1$. For $t \in (0, 1)$, the sequence x_t is partially masked, with each being masked with probability α_t (*i.e.,* a noise schedule) or remaining unmasked with probability $1 - \alpha_t$. The reverse process recovers the data distribution by iteratively predicting masked tokens at t moves from 1 to 0.

To train a MDLM, a forward process with a specific form of α_t is designed. We parameterize a bi-directional unmasking predictor $f_\theta(\cdot|x_t)$ that takes x_t as input and predicts all masked tokens simultaneously. In this study, we highlight the specific noise schedule $\alpha_t = t$ designed in LLaDA and d1 (Zhao et al., 2025), which is a representative of MDLMs. Then the objective function of them is presented with the cross-entropy loss computed only on the masked tokens:

$$\mathcal{L}(\theta) = -\mathbb{E}_{t,\, x_0,\, x_t}\left[\frac{1}{t\,|x_t|}\sum_{k\in\mathcal{M}_t}\log f_\theta\Big(x_0^k \mid x_t\Big)\right],$$

$$\mathcal{M}_t = \Big\{k \;\Big|\; x_t^k = \texttt{[MASK]}\Big\}$$

where t is sampled from an uniform distribution $t \sim \mathcal{U}(0, 1)$, $x_0 \sim p_{\text{data}}$ is the observable data collected from real world, $x_t \sim q_{t|0}(x_t \mid x_0)$ is a fully masked sequence applied by α_t and $\mathcal{M}_t$ denotes the set of positions of the masked tokens. Note that the loss is only calculated for tokens that are masked out in timestep t.

It is worth noting that the key difference between MDLMs and BERT (Devlin et al., 2019) is that the latter uses a fixed masking ratio and the decoding is a single-step infilling process, whereas MDLMs use time-varying noise schedule for masking and the decoding process performs multi-step denoising from the full noise.

4 Our Method: Diff4TST

Our approach formulates text style transfer as a conditional generation problem under an MDLM framework. As illustrated in Figure 1, we adopt a diffusion language model to progressively transform a source style sentence into the target style through a sequence of denoising steps.

To support effective stylistic control, we design a style-aware noise schedule that allows the model to selectively perturb stylistic components while retaining non-stylistic or content information. At the SFT stage, for each paired example, we extract the token-level edits by aligning the source to target text, allowing the model to explicitly distinguish stylistic modification from content preservation. The MDLM is fine-tuned on these stylistic contents, encouraging it to rewrite stylistic tokens while copying content tokens, thereby learning to perform controlled style transfer within the diffusion process (cf. Section 4.1).

At the inference stage, given a source sentence and a desired target style, we first generate an initial rewrite using the fine-tuned diffusion language model. To ensure that the generated sentence adheres to the target style without additional training, we introduce a generate-then-refine inference framework. After obtaining the initial output, we employ an external style classifier to compute token-level gradient attribution with respect to the target style. Tokens that contribute most to the style mismatch are identified and selectively re-masked, after which the diffusion language model performs an additional denoising step to refine the output (cf. Section 4.2).

4.1 Diffusion Language Model for TST

4.1.1 Unmasking Predictor for TST $f_\theta(\cdot|x_t)$

We define the unmasking predictor used for TST task as a conditional text generator, where x_0 consists of an instruction followed by the source text and target text, denoting as:

$$x_0 = \texttt{PROMPT} \oplus X \oplus Y \oplus \texttt{[EOS]},$$

where PROMPT is the instruction template indicating the target style. X and Y are the parallel text sampled from the training set. [EOS] is the marker of end-of-answer.

Considering the task formulation, we prompt the model to condition on the instruction template and source text, generate the target text. Hence, x_t is in the format of:

$$x_t = \texttt{PROMPT} \oplus X \oplus \texttt{[MASK]}, ..., \texttt{[MASK]} \oplus \texttt{[EOS]}. \quad (1)$$

Our unmasking predictor $f_\theta(\cdot|x_t)$ is able to infer the masked tokens in Y based on its contexts. We follow the model design of LLaDA (Nie et al.) and adopt the open-source backbone *GSAI-ML/LLaDA-8B-Instruct* as the unmasking predictor. Its architecture is specifically designed for conditional token restoration under partial corruption, making it well-suited to our unmasking objective.

4.1.2 Style-aware Noise Schedule α_t

The noise schedule α_t plays a central role in diffusion-based text generation, as it governs the formation of the style-aware masked token set $\mathcal{M}_t$ at timestep t. The existing MDLMs usually consider the same probability for all the tokens in x_0 during masking. In LLaDA, the noise schedule $\alpha_t = t$ specifies that each token is independently masked with unified probability t during supervised fine-tuning. However, this design is suboptimal for text style transfer, as a uniform noise schedule is fundamentally misaligned with the sparse-edit nature of the task. By distributing learning signals uniformly across all tokens, it fails to emphasize stylistically salient regions that actually require modification. Moreover, applying denoising to largely immutable or factual tokens introduces unnecessary training noise, hindering the model's capacity for minimal and precise stylistic refinement. Together, these limitations motivate a noise schedule that explicitly differentiates stylistic edits from content preservation.

To bridge the gaps, we design a style-aware noise schedule. The motivation for this design stems from the observation that stylistic tokens are frequently edited during style transfer, thereby making them easier to identify, whereas non-stylistic tokens are typically preserved. Therefore, we first identify the type of tokens in text. For each (X, Y) pair, we apply a token-level Levenshtein alignment algorithm that dynamically selects the alignment maximizing matches between the source and target texts, and extracts token-level edit operations (Bryant et al., 2023). This process yields a sequence of token-level annotations that distinguish unchanged tokens ([COPY]) from edited tokens ([EDIT]), including insertions, deletions, and substitutions.

During the forward process, we sample a timestep $t \in (0, 1)$ and corrupt tokens in Y to obtain a noise text x_t, where the noise schedule α_t for [COPY] and [EDIT] tokens are defined as:

$$\alpha_t(\texttt{[EDIT]}) = t,$$
$$\alpha_t(\texttt{[COPY]}) = (1 - r)\,\frac{1 - \cos(\pi t)}{2},$$

where r denotes the edit ratio of the text, *i.e.,* the fraction of the target tokens labeled as [EDIT]. Compared to the linear noise schedule $\alpha_t = t$, the flat initial slope of the cosine-based schedule avoids excessive corruption in early timesteps, while its steeper behavior in later timesteps enables stronger refinement. A larger r suppresses the copy rate, encouraging the model to focus more on the edited spans rather than redundant token preservation. Together, this noise schedule design induces a masked token set $\mathcal{M}_t$ that biases the diffusion process toward stylistic refinement.

Besides, we also incorporate different weights for [COPY] and [EDIT] tokens by introducing a

style-aware coefficient β_t to the cross-entropy loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{t,x_0,x_t}\left[\frac{1}{t\,|x_t|}\sum_{k\in\mathcal{M}_t}\beta_t^k \log f_\theta\Big(x_0^k \mid x_t\Big)\right],$$

$$\beta_t^k = \begin{cases} t, & x_0^k \quad \text{with [EDIT]}, \\ 1-t, & x_0^k \quad \text{with [COPY]}. \end{cases}$$

Accordingly, we design the noise schedule to primarily focus on edited tokens, encouraging the model to learn how to substitute stylistic tokens under minimal modification constraints. This probability reflects the intrinsic transformation of text style transfer and leads to more controllable and semantically faithful generation.

4.2 Generate-then-refine Inference

Unlike the traditional reverse process of MDLMs, we design a generate-then-refine process to denoise the text. During the inference stage, our method first *generates* a complete text, which serves as a strong initial hypothesis for style transferring. Then, our method *refines* the hypothesis by revisiting the stylistic tokens in it and partially regenerates these tokens.

Generation stage. We follow the reverse denoising process of MDLMs and iteratively denoise the masked sequence. We initialize x_t as a fully masked target sequence, as shown in Equation 1. At each reverse step, the unmasking predictor $f_\theta(\cdot \mid x_t)$ takes the current partially masked sequence as input and predicts token distributions for masked positions in parallel. Following LLaDA, we adopt a low-confidence re-masking strategy, where tokens with low prediction confidence are re-masked. Formally, for each masked position k, we sample

$$x_{t-1}^k \sim \text{Categorical}\big(f_\theta(\cdot \mid x_t)_k\big).$$

where x_{t-1}^k denotes the token sampled at position k at the next reverse step, which ensures the transition of the reverse process aligns with the forward process for accurate sampling. Through iterative prediction and confidence-based re-masking, the sequence gradually converges to a complete and stable output $\tilde{x}_0$ at this stage.

Refinement stage. Regarding $\tilde{x}_0$, there may exist some attribute discrepancies. For example, given the task "rewrite the text for elementary students", the college level input sentence is "Education is the most potent and intellectually consequential instrument which one may employ to effect transformative change in the global milieu." is expected to be simplified to the elementary level. However, after generation stage, the model outputs "Education is the powerful weapon that you can use to change the world.", which is middle level rather than the desired elementary level. This discrepancy motivates the refinement stage, where gradient-based token re-masking selectively identifies high-complexity terms for regeneration. To omit the discrepancies, we re-mask the stylistic tokens in the generated hypothesis via gradient analysis following past work (Ebrahimi et al., 2018). We define a set of the re-masked tokens as:

$$\mathcal{M} \leftarrow \text{argmax}_{\tilde{x}_0^i \in \tilde{x}_0} \|\nabla_{\tilde{\mathbf{x}}_0^i} \mathcal{L}_{\text{classifier}}\|_2,$$

where the cross-entropy loss $\mathcal{L}_{\text{classifier}}$ was computed based on a style classifier based on `distilbert-base-uncased` with a single linear classification head. The lightweight style classifier was trained using the same training data as Diff4TST. Let $\tilde{\mathbf{x}}_0^i$ denote the embedding of the i-th token. We take the ℓ_2-norm of the gradient $\|\nabla_{\tilde{\mathbf{x}}_0^i} \mathcal{L}_{\text{classifier}}\|_2$ as its importance score and rank all tokens accordingly. To perform selective refinement, we re-mask the top tokens based on their gradient magnitude, which is adaptively determined by the distance to the target style.

We re-perform denoising but on re-masked tokens in $\tilde{x}_0$ by conducting the same sampling as mentioned in generation stage. After a full sequence is predicted, we evaluate whether the refinement meets the target level. If satisfied, the refinement loop terminates early; otherwise, the gradient-based refinement and denoising are repeated until convergence or until a predefined maximum iteration budget is reached. This results in the final generation outcome $\hat{x}_0$.

5 Experimental Setup

5.1 Datasets and Evaluation

We evaluate Diff4TST on representative text style transfer benchmarks with varying granularity.

Polarity style. The target style is binary, and the task requires transforming text between two opposing attributes. We evaluate our model on two widely used public benchmarks to facilitate comparison with prior work. **GYAFC** (Rao and Tetreault, 2018) is a sentence-level formality transfer dataset consisting of parallel {*informal* $\leftrightarrow$ *for-*

Model	Style Transfer Accuracy ↑				Content Preservation (LaBSE) ↑				Fluency (Perplexity) ↓			
	GYAFC		ParaDetox		GYAFC		ParaDetox		GYAFC		ParaDetox	
	informal ⇔ formal		toxic ⇔ neutral		informal ⇔ formal		toxic ⇔ neutral		informal ⇔ formal		toxic ⇔ neutral	
LLaMA-3-8B	80.00	11.20	47.67	29.04	0.75	**0.90**	0.77	0.86	92.53	87.69	113.84	191.30
APE	74.00	12.20	47.57	28.44	0.75	0.88	0.78	0.87	94.27	89.93	133.12	188.34
AVF	76.00	12.40	47.57	28.44	0.74	0.88	0.78	0.87	96.63	89.36	131.10	191.29
PNMA	73.85	8.70	42.43	23.79	**0.79**	0.90	0.79	**0.89**	103.61	90.85	136.27	194.71
sNeuron-TST	80.80	14.40	55.36	31.98	0.74	0.89	0.74	0.81	90.79	**81.46**	85.65	172.26
Diff4TST	**99.12**	**84.56**	**98.36**	**96.34**	0.72	0.77	**0.81**	0.86	**33.33**	90.36	**63.18**	**89.08**

Table 1: Main results of polarity style transfer. "Style Transfer Accuracy" measures the accuracy of the labels predicted by a well-trained style classifier. "Content Preservation" measures the preserved contents of the original text. "Fluency" measures the perplexity of the generated sentences. The baseline results are copied from study (Lai et al., 2024).

mal} sentence pairs collected from Yahoo Answers. Following standard practice, we use the *Family & Relationships* domain for evaluation. **PARADETOX** (Logacheva et al., 2022) is a dataset for toxicity style transfer with polarity {*toxic* ↔ *neutral*}, where toxic sentences are rewritten into neutral ones with keeping semantics unchanged.

Fine-grained style. The target style contains multiple intensities. **CNN/DM** (Hermann et al., 2015) is a fine-grained readability transfer dataset. Following prior work, we compute the Flesch Reading Ease (FRE) score (Flesch, 1948) for both the original texts and their summaries, and partition the data into four readability levels: {*elementary school (1)* → *middle school* (2) → *high school* (3) → *college* (4)}. **YELP** is a review dataset (Zhang et al., 2015) for fine-grained sentiment transfer. Each review is labeled with a rating from 1 to 5 stars, corresponding to {*very negative (1)* → *negative* (2) → *neutral* (3) → *positive* (4) → *very positive* (5)}.[1]

We evaluate our approach with official evaluation metrics for these benchmarks following standard evaluation protocols[2].

5.2 Baselines

Polarity style. For polarity-based style transfer, we compare our method against a set of *autoregressive* baselines that steer generation by manipulating internal neuron activations: (1) **LLaMA-3-8B**, used without additional fine-tuning; (2) **APE** (Tang et al., 2024); (3) **AVF** (Tan et al., 2024b); (4) **PNMA** (Kojima et al., 2024); and (5) **sNeuron-TST** (Lai et al., 2024).

Fine-grained style. We compare our diffusion-based approach with baselines built upon autoregressive language modeling and reinforcement learning. Specifically, we include: (1) **GPT-4o-mini** and **GPT-5** as autoregressive baselines and (2) **LLaDA-8B-Instruct** as zero-shot where style transfer is achieved purely via prompting without task-specific fine-tuning. In addition, we consider (3) an **SFT+PPO** baseline based on a **T5-large** backbone (Gu et al., 2026) , following the standard supervised fine-tuning followed by reinforcement learning optimization paradigm for controllable text generation. This baseline relies on carefully designed reward functions to encourage style alignment while preserving semantic content.

5.3 Implementation Details

We follow (Zhao et al., 2025) for implementations and training configurations of diffusion language models. The model is trained with AdamW (1×10^{-5} lr, 0.1 weight decay) for 10 epochs with global batch size 16, FP16 precision, and gradient clipping at 1.0. Generation uses 64 diffusion steps, and refinement is limited to at most 5 iterations.

6 Results

6.1 Main Results

As shown in Table 1 and Table 2, we evaluate Diff4TST on both fine-grained and polarity style transfer benchmarks. Overall, our method achieves lower deviation from target styles than all baselines by an overwhelming majority, indicating the superiority of our method in various text style transfer scenarios.

Regarding polarity style transfer tasks, Diff4TST attains high style transfer accuracy in arbitrary directions. As shown in Table 1, our method improves formality transfer accuracy on GYAFC from state-of-the-art 80.8/14.4 (sNeuron-TST) to 99.1/84.6, and achieves over 95% accuracy on both directions of toxicity transfer on ParaDetox. At the same time, Diff4TST maintains comparable or better content preservation and achieves lower per-

[1]The details of the datasets are shown in the Appendix A

[2]We display the detailed evaluation metrics for each benchmark in Appendix B.

Readability (CNN/DM)										
Level (s_y)	**GPT-5**		**GPT-4o-mini**		**LLaDA-8B-Instruct**		**T5-large (SFT+PPO)**		**Diff4TST**	
	FRE	**FRE$_\Delta$** ↓	**FRE**	**FRE$_\Delta$** ↓	**FRE**	**FRE$_\Delta$** ↓	**FRE**	**FRE$_\Delta$** ↓	**FRE**	**FRE$_\Delta$** ↓
Elementary school	75.02	14.98	70.96	19.04	71.12	18.88	79.96	10.04	**85.42**	**4.58**
Middle school	67.66	2.34	65.41	4.59	68.01	1.91	64.65	5.35	**69.36**	**0.64**
High school	58.65	8.65	60.73	10.73	60.77	10.77	48.92	1.08	**50.10**	**0.10**
College	45.33	25.33	47.42	27.42	45.59	25.59	25.71	5.71	**24.28**	**4.28**
Average	–	12.83	–	15.44	–	14.28	–	5.55	–	**2.40**

Sentiment (Yelp)										
Level (s_y)	**GPT-5**		**GPT-4o-mini**		**LLaDA-8B-Instruct**		**T5-large (SFT+PPO)**		**Diff4TST**	
	STAR	**STAR$_\Delta$** ↓	**STAR**	**STAR$_\Delta$** ↓	**STAR**	**STAR$_\Delta$** ↓	**STAR**	**STAR$_\Delta$** ↓	**STAR**	**STAR$_\Delta$** ↓
Very Negative	1.2802	0.2802	1.5190	0.5190	1.3066	0.3066	1.2114	0.2114	**1.0186**	**0.0186**
Negative	1.1210	0.8790	1.8018	0.1982	1.3156	0.6844	2.3234	0.3234	**2.0940**	**0.0940**
Neutral	2.3099	0.6901	2.9476	0.0524	2.2315	0.7685	3.3748	0.3748	**2.9990**	**0.0010**
Positive	4.2219	0.2219	4.3652	0.3652	3.8888	**0.1112**	4.1462	0.1462	**4.3084**	0.3084
Very Positive	4.9533	0.0467	4.6032	0.3968	4.8249	0.1751	4.9374	0.0374	**5.0000**	**0.0000**
Average	–	0.4236	–	0.3063	–	0.4092	–	0.2186	–	**0.0844**

Table 2: Main results of fine-grained style transfer. "FRE" measures the readability score of the text. "STAR" denotes the sentiment intensity of the text predicted by a well-trained classifier. We also report target attribute scores and deviation from desired styles (Δ, lower is better). The baseline results are copied from study (Gu et al., 2026).

Readability Level	**Target Flesch**	**Method**	**Flesch Score**	**FRE$_\Delta$** ↓	**Success Rate** ↑
Elementary	90	w\o style-aware noise schedule	79.97	10.03	49.41%
		w\o generate-then-refine framework	83.61	6.39	70.44%
		Diff4TST	**85.42**	**4.58**	**92.36%**
Middle	70	w\o style-aware noise schedule	65.86	14.14	57.64%
		w\o generate-then-refine framework	67.68	2.32	66.98%
		Diff4TST	**69.36**	**0.64**	**93.91%**
High	50	w\o style-aware noise schedule	50.55	0.55	45.63%
		w\o generate-then-refine framework	49.41	0.59	56.31%
		Diff4TST	**50.10**	**0.10**	**92.04%**
College	20	w\o style-aware noise schedule	31.10	11.10	77.32%
		w\o generate-then-refine framework	24.68	4.68	85.80%
		Diff4TST	**24.28**	**4.28**	**96.88%**
Overall	–	w\o style-aware noise schedule	–	8.96	57.35%
		w\o generate-then-refine framework	–	3.50	69.77%
		Diff4TST	–	**2.40**	**93.77%**

Table 3: Ablation study on CNN/DM benchmark. "Target Flesch" denotes the target score of different levels. "Success Rate" denotes the proportion of samples that reach the target level.

plexity on most cases, indicating the outstanding capability in text generation of our method. These results demonstrate that our MDLM framework can effectively balance style control, content preservation, and fluency under polarity constraints.

Regarding fine-grained style transfer tasks, Diff4TST yields substantially smaller attribute deviations across all target levels. Especially, on CNN/DailyMail benchmark, our method reduces the average FRE deviation to 2.40, whereas strong large language model baselines operating in a zero-shot manner exhibit deviations exceeding 14 on average. Similarly, on Yelp benchmark, Diff4TST achieves the lowest average STAR deviation (0.084), indicating much stronger alignment with the desired sentiment levels. These results suggest that our method is particularly effective in intensive style space that requires precise control over the style intensity levels.

6.2 Ablation Study

We conduct ablation studies on CNN/DM benchmark to analyze how different components contribute to readability style transfer.

Ablation on framework components. As we can see from Table 3, when adopting LLaDA's original noise schedule $\alpha_t = t$, the model applies a uniform probability across all tokens. The dramatic performance depreciation confirms that uniform masking disperses attention across the entire sentence and fails to emphasize style-bearing pivot tokens. Discarding the generate-then-refine inference strategy also degrades the performance. But compared with the effect of the style-aware noise schedule, it is less significant. The results suggest that the style-aware noise schedule and the generate-then-refine inference play complementary roles in achieving precise control of readability.

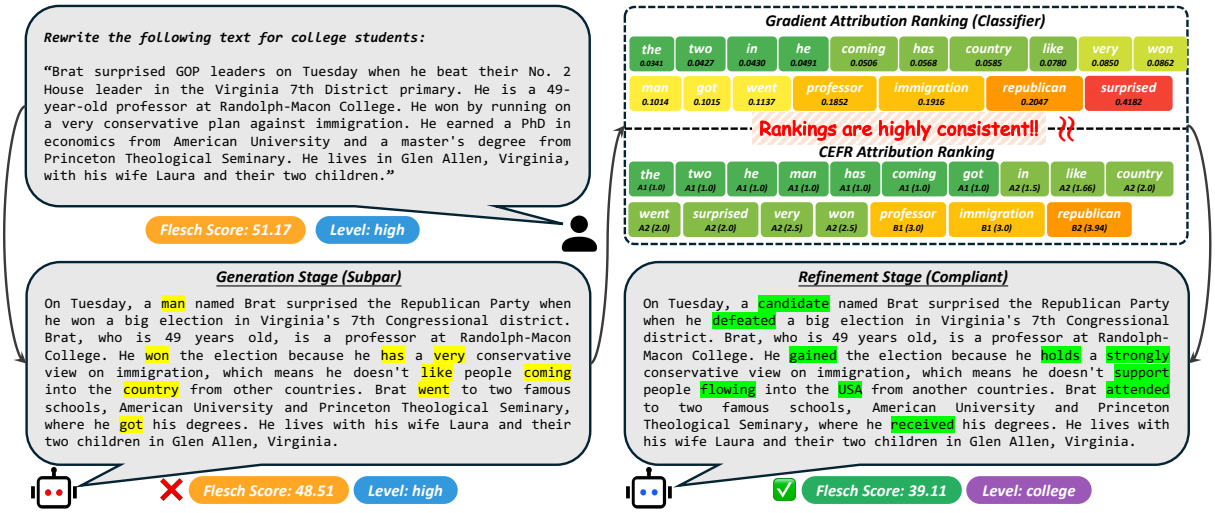

Figure 2: An example showcasing the effect of generate-then-refinement inference. The text produced at the generation stage cannot satisfy the target style. Next, stylistic tokens are re-masked (shown with yellow highlight in the left bottom box) and re-generated (shown with green highlight in the right bottom box). We display the identified attributed tokens in the right top box, which exhibit a high degree of consistency with CEFR levels.

We further visualize the difficulty transitions across target readability levels to examine the stability of fine-grained control.[3]

Ablation on noise schedule. As shown in Table 4, we further study the impact of different copy-mask scheduling strategies on readability style transfer. Results show that dynamically adjusting the copy probability consistently improves controllability compared to a fixed noise schedule, leading to lower attribute deviation and higher target hit rates. [4]

Readability Level	Method	Flesch Score	FRE$_\Delta$ ↓	Success Rate ↑
Elementary	Fixed NS	79.06	10.94	47.00%
	Diff4TST*	**83.61**	**6.39**	**70.44%**
Middle	Fixed NS	66.07	3.93	62.10%
	Diff4TST*	**67.68**	**2.32**	**66.98%**
High	Fixed NS	**49.73**	**0.27**	47.80%
	Diff4TST*	49.41	0.59	**56.31%**
College	Fixed NS	27.11	7.11	85.00%
	Diff4TST*	**24.68**	**4.68**	**85.80%**
Overall	Fixed NS	–	5.56	60.48%
	Diff4TST*	–	**3.50**	**69.77%**

Table 4: Ablation study on CNN/DM benchmark. "Fixed NS" denotes we fix **N**oise **S**chedule in the diffusion model for [COPY] tokens, which is a fixed value of 0.3, indicating the overall ratio of copy labels in the training corpus rather than fluctuating with each sentence. Diff4TST* denotes the proposed model trained with the style-aware cosine noise schedule, without the generate-then-refine inference stage.

[3]Additional analysis and visualization of readability-level transitions are provided in Appendix C.

[4]Detailed comparisons and analysis are provided in Appendix D.

6.3 Case Study

Figure 2 presents an example of CNN/DM. The original sentence is identified as the *high* level with a Flesch score of 51.17, and the task requires to transfer it to the *college* level with a target Flesch score of 20. Directly applying the fine-tuned diffusion language model leads to a sentence with reduced Flesch score but still does not attain the target range. After applying the refinement stage, these tokens are selectively re-masked and regenerated with full contextual access, encouraging the model to modify them toward a target style. The final output successfully falls within the target Flesch score range. Notably, the tokens selected by our gradient analysis closely aligns with the CEFR-based (Common European Framework of Reference for Languages) (Little, 2006) vocabulary readability assessment, further validating the effectiveness of our token-level refinement strategy.

7 Conclusion

In this paper, we propose Diff4TST, a diffusion-based method for text style transfer, which enables targeted stylistic transfer while preserving the semantic content of a given text. Diff4TST incorporates a style-aware noise schedule that prioritizes stylistic pivots over uniform token corruption and a generate-then-refine inference framework that facilitates controllable rewriting. Extensive experiments on four text style transfer benchmarks demonstrate the superiority of Diff4TST.

Limitations

Despite its effectiveness, Diff4TST has several limitations. First, the generate-then-refine inference procedure introduces additional iterative steps compared to single-pass autoregressive decoding, which may lead to increased latency in real-time or low-latency settings. Moreover, our experiments are limited to English datasets and sentence-level rewriting. Extending Diff4TST to multilingual scenarios, longer-form documents, and more complex discourse-level style transformations remains an important direction for future work.

Acknowledgments

This work was supported in part by the National Key Research & Develop Plan (Project No.2023YFF0725100) and Natural Science Foundation of China (Project No. U23A20298), and in part by the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Project No.24CGA26).

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Rudolf Franz Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *The Thirteenth International Conference on Learning Representations*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Diffuseq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations (ICLR 2023)(01/05/2023-05/05/2023, Kigali, Rwanda)*.

Shuhuan Gu, Wenbiao Tao, Xinchen Ma, Kangkang He, Ye Guo, Xiang Li, and Yunshi Lan. 2026. Unsupervised text style transfer for controllable intensity. *Preprint*, arXiv:2601.01060.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Jingxuan Han, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15187–15201, Bangkok, Thailand. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971. Association for Computational Linguistics.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Joosung Lee. 2020. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 195–204.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018b. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Xiaolin Li, Lei Huang, Yifan Zhou, and Changcheng Shao. 2021. Tst-gan: A legal document generation model based on text style transfer. In *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, pages 90–93. IEEE.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning*, pages 21051–21064. PMLR.

David Little. 2006. The common european framework of reference for languages: Content, purpose, origin, reception and impact. *Language Teaching*, 39(3):167–190.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Sourabrata Mukherjee and Ondřej Dušek. 2023. Leveraging low-resource parallel data for text style transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander M Rush, Thomas PIERROT, and Volodymyr Kuleshov. 2025. Simple guidance mechanisms for discrete diffusion models. In *The Thirteenth International Conference on Learning Representations*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.

Keren Tan, Kangyang Luo, Yunshi Lan, Zheng Yuan, and Jinlong Shu. 2024a. An llm-enhanced adversarial editing system for lexical simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1136–1146.

Shaomu Tan, Di Wu, and Christof Monz. 2024b. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Andrew Zhang, Anushka Sivakumar, Chia-Wei Tang, and Chris Thomas. 2025. Flexible-length text infilling for discrete diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31332–31347, Suzhou, China. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. 2025. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *Preprint*, arXiv:2504.12216.

Kun Zhou, Yifan Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1438–1451.

Chang Zong, Yuyan Chen, Weiming Lu, Jian Shao, Yongfeng Huang, Heng Chang, and Yueting Zhuang. 2024. Proswitch: Knowledge-guided instruction tuning to switch between professional and non-professional responses. *arXiv preprint arXiv:2403.09131*.

A Datasets Details

Table 5 reports the statistics of the constructed training datasets for fine-grained style transfer. All datasets are used exclusively for supervised fine-tuning, where parallel sentence pairs are formed by aligning samples across different style attributes.

Dataset	Scores	Attributes	Train	Test
CNN/DailyMail	$80 \leq \text{FRE} \leq 100$	Elementary School	4,000	4,000
	$60 \leq \text{FRE} < 80$	Middle School	4,000	
	$40 \leq \text{FRE} < 60$	High School	4,000	
	$0 \leq \text{FRE} < 40$	College	4,000	
Yelp	5 stars	Very Positive	4,000	5,000
	4 stars	Positive	4,000	
	3 stars	Neutral	4,000	
	2 stars	Negative	4,000	
	1 star	Very Negative	4,000	

Table 5: Overview of fine-grained style attributes used in our experiments. For readability control, samples are grouped by Flesch score ranges on CNN/DailyMail. For sentiment control, Yelp reviews are categorized by star ratings. The training size denotes the number of constructed parallel sentence pairs per attribute level.

B Evaluation Metrics

B.1 Polarity style

Style Accuracy. We report the accuracy of labels predicted as correct by a style classifier.

Content Preservation. We measure content preservation by computing the cosine similarity between the embeddings of the original text and the text generated by the model. Sentence embeddings are obtained using LaBSE (Feng et al., 2022) as our primary metric.

Fluency. We evaluate fluency using the perplexity of the generated sentences computed by GPT-2 (Radford et al., 2019).

B.2 Fine-grained style

Readability. We measure readability using the *Flesch Reading Ease* (FRE) score, a widely adopted metric that reflects sentence complexity and word-level difficulty. For fine-grained control, we quantify deviation from the target readability level by computing the absolute difference between the FRE score of the generated text and the midpoint of the target readability range. The *Flesch Reading Ease* (FRE) score to quantify the readability of generated text is defined as:

$$\text{FRE} = 206.835 - 1.015 \cdot \frac{\text{total words}}{\text{total sentences}} - 84.6 \cdot \frac{\text{total syllables}}{\text{total words}}. \quad (2)$$

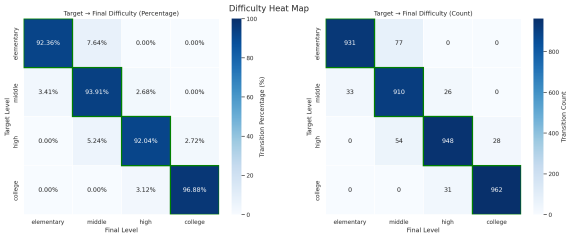

Figure 3: Difficulty transition heatmap showing generation results across four target readability levels.

To measure deviation from the target readability level, we compute:

$$\text{FRE}_\Delta = \left|\text{FRE}_{\hat{y}} - \text{FRE}_{s_y}\right|, \quad (3)$$

where $\text{FRE}_{\hat{y}}$ denotes the FRE score of the generated text, and FRE_{s_y} is the midpoint of the predefined target readability range.

Sentiment. For sentiment transfer, we employ a pretrained sentiment classifier to predict the sentiment intensity (*STAR*) of the generated text. We measure controllability by computing the absolute deviation between the predicted sentiment class and the target sentiment label, which is defined as:

$$\text{STAR}_\Delta = \left|\text{STAR}_{\hat{y}} - \text{STAR}_{s_y}\right|, \quad (4)$$

where $\text{STAR}_{\hat{y}}$ is the predicted sentiment class of the generated text, and STAR_{s_y} is the target sentiment label.

C Additional Analysis on Readability Control

As shown in Figure 3, the difficulty transition heatmap presents a clear diagonal dominance across the four readability targets (*elementary–college*), suggesting that generated texts strongly align with the desired difficulty levels. Our approach achieves more than **92%** accuracy for all transitions, with very limited off-target drift, a pattern that also persists across other evaluation datasets.

D Additional Ablation on Copy-Mask Scheduling

We investigate how different copy-mask scheduling strategies influence style transfer performance during the supervised fine-tuning (SFT) stage. Readability transfer typically requires modifying only a small subset of stylistic tokens, while most factual content should be preserved. Under this setting,

uniformly masking COPY tokens may introduce unnecessary noise and limit the model's ability to perform subtle stylistic refinement. We compare two copy-mask strategies. (1) **Fixed Noise Schedule**, where COPY tokens follow a constant masking ratio, and (2) **Cosine Noise Schedule**, where the copy probability is dynamically annealed using a cosine schedule. The cosine strategy consistently achieves lower attribute deviation and higher target hit rates across all readability levels. This indicates that dynamically adjusting copy probabilities encourages balanced editing behavior, preventing over-rewriting while better preserving factual content.